# Summary Report

Researcher: Yulong Gong

Nowadays, almost every company will have some data related positions. Big data is the theme of this era and more and more managers realize the power of data, especially those from their own business. Not only data can help them keep track of their business performance, but also it can help them evaluate their marketing strategies, especially product design and customer clustering. For this project, I choose a Kaggle dataset of an automobile company who wants to enter a new market.

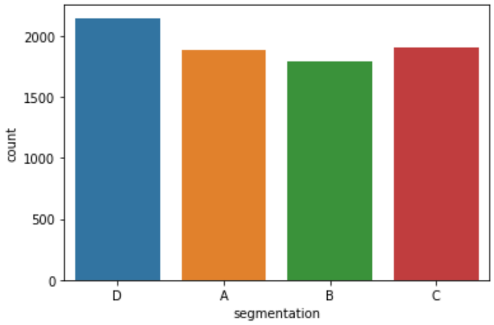
The dataset contains a training set and test set, for the purpose of the project, I would focus on the training set only. The training set is composed of 8068 observations, each observation has 11 features, which includes gender,  marital status, profession, spending scores, etc. 6 out of 11 columns containing missing values, which are ever_married, graduated, profession, work_experience, family_size and var_1 (an anonymised category defined by the company).

For the family size, it is reasonable to replace the missing values with the median, since there are few large outliers which may affect the mean value, so the mean value is not representative of the population. Even though there are few profession categories, I still consider it to be individual specific, so it is hard to get the missing values for profession and I will drop those records.  As far as I'm concerned, profession is related to the graduate and work experience. To be specific, for an individual to have a profession, he/she should graduate. But if an individual has a profession, it doesn't mean he/she must have work experience. So if an individual has a profession, the missing graduate status will be filled with yes, and missing work experience would be filled with the minimum, which is 0 in this case. For the marital status and var1, observations with missing columns are dropped.

Furthermore, the dataset contains numeric columns and string columns. For the numeric columns, data are recorded at different scales. To make each column have a similar weight in the prediction, standardization is necessary. For string columns, dummif is performed to maintain the information captured by strings. After the data preprocessing, I have 7736 rows with 22 columns. To fully understand the data, the next step is Exploratory data analysis is the way to go.

As shown in figure 1, it can be suggested that the company classified their customers into four groups, group D has the most consumers and group B has the least consumers. However, the number difference is small and I would see the group is divided relatively evenly. So using the accuracy matrix to mature prediction accuracy will be reasonable.

*Figure 1. Segmentation distribution*



As shown in Figure2 the dataset is distributed relatively evenly gender wise. There are more male than females. Surprisingly, as indicated in Figure 3, there are more number of male has a career in healthcare than females. And in the fields of engineer, there are more females than males. Since the number of record differences in gender is not so significant, it is less likely the gender difference would make such a big difference in prediction.

*Figure 2. Gender distribution*

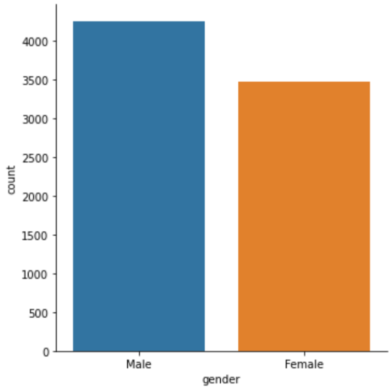
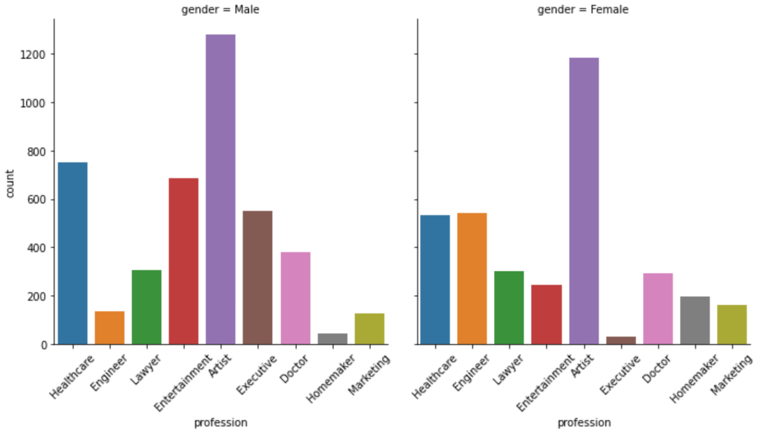
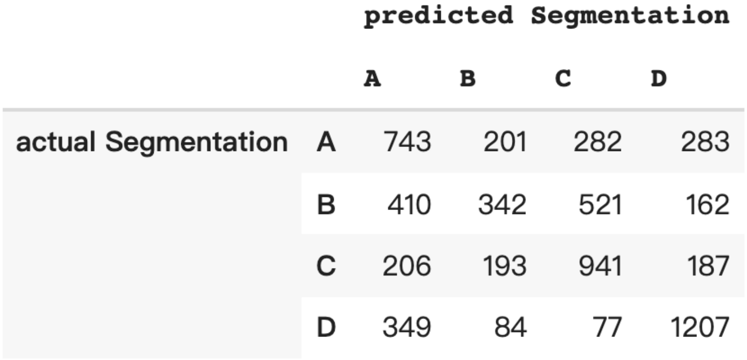


*Figure 3. Profession distribution by gender*



After having a better understanding of the dataset, especially to decide which matrix I would use to evaluate the model and whether there are potential issues needed to be considered such as race and gender, I decided to try to build models to predict which segmentation a customer belongs to, which is a problem the company trying to solve. Since I would not use the test set in this project, I would split the training set into training(0.8) and validation(0.2) to check the model performance and whether an overfitting or underfitting issue exists. I start with logistic regression whose accuracy is 0.52, which is not so different from flipping a coin. As shown in Chart 1, this model did an ok job in terms of predicting those who belong to segmentation D. But it cannot capture those who belong to segmentation A.

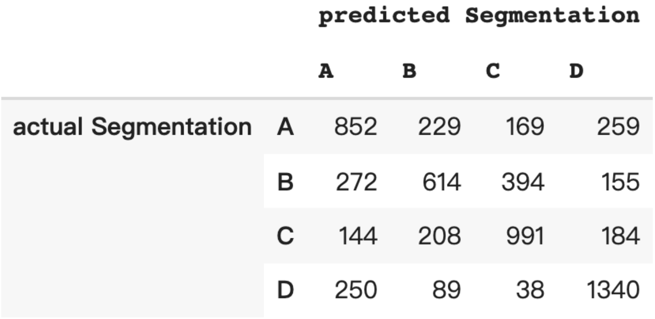
*Chart1. Confusion matrix of Logistic Regression*



| | | predicted Segmentation | | | |
|---|---|---|---|---|---|
| | | A | B | C | D |
| actual Segmentation | A | 743 | 201 | 282 | 283 |
| | B | 410 | 342 | 521 | 162 |
| | C | 206 | 193 | 941 | 187 |
| | D | 349 | 84 | 77 | 1207 |

Then I decided to try another model, which is xgboost, for the xgboost I tried some other parameters rather then default, which has an accuracy of 61% in the validation set. The concussion matrix is shown in chart 2. Even though the model performance increased by 10%, there is still some false prediction.

*Chart2. Confusion matrix of xgboost.*

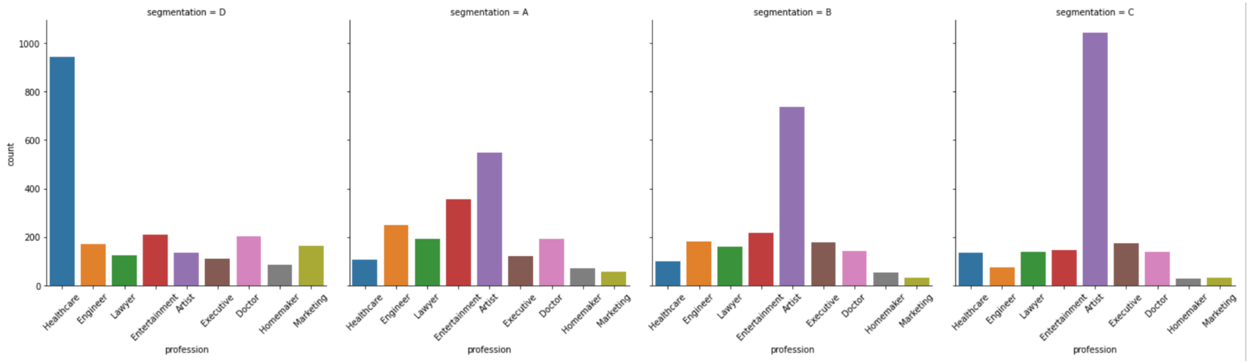


| | | predicted Segmentation | | | |
|---|---|---|---|---|---|
| | | A | B | C | D |
| actual Segmentation | A | 852 | 229 | 169 | 259 |
| | B | 272 | 614 | 394 | 155 |
| | C | 144 | 208 | 991 | 184 |
| | D | 250 | 89 | 38 | 1340 |

.

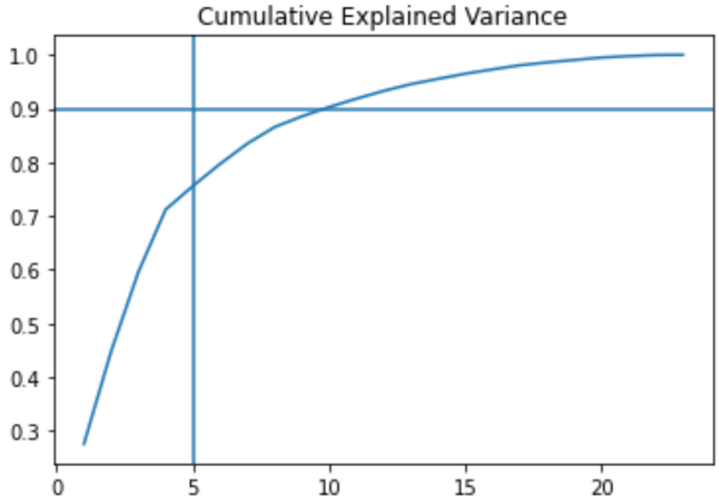
Logistic is a classic model in classification and xgboost is powerful, but they didn't get a good performance as I expected. So I think based on the data provided, it is hard to classify each individual into appropriate groups. So I checked the profession by segmentation and surprisingly, as shown in Figure 4, group D has the most healthcare individuals, and group C has the most number of artists. However, for group B and group A, the difference is small, which may indicate that the company decided the number of products not based on their marketing segmentation. So I would like to use unsupervised learning techniques whether the company classifies their customers based on the results shown by their dataset.

*Figure4. Profession distribution by segmentation*



After the data preprocessing, I got 22 columns. Since unsupervised learning is based on the distance between datapoints. And in high dimension, the distance becomes hard to interpret, so dimension reduction technique is necessary. To decide how many components to keep, I checked the explained variance and cumulative explained variance plot as shown in Figure 5.From the cumulative explained variance chart, we can conclude that around 10 components would explain 90% of the variance in the dataset. However, when we use only the 5 components, we can only explain 70% of the variance. I would choose only the first 10 components because I was trying to find a balance between reducing dimensions and explaining as much variance as possible.

*Figure 5. Cumulative Explained Variance*



It also becomes a question of how many clusters I would choose. So I tried a set of potential cluster numbers as shown in Figure 7, there is no elbow in the inertia plot, but in the silhouette score plot, 3 clusters actually have the highest silhouette score. As shown in the silhouette plot in Figure 6, the silhouette score is 0.231, and there is no negative value, which indicates 3 clusters are actually a good cluster of the current data we have. And from the same plot, we can see that cluster 1 has the most number of people, and group 0 has the least number of people.

*Figure 6. Silhouette plot*

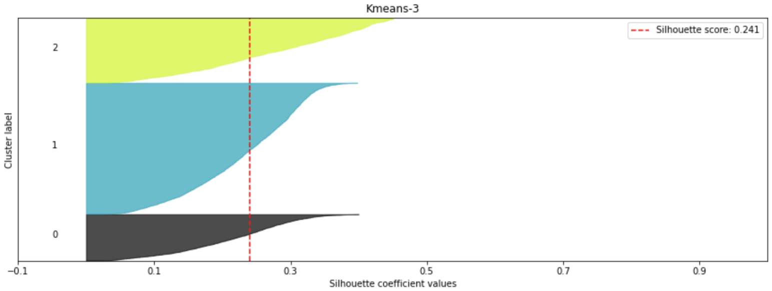
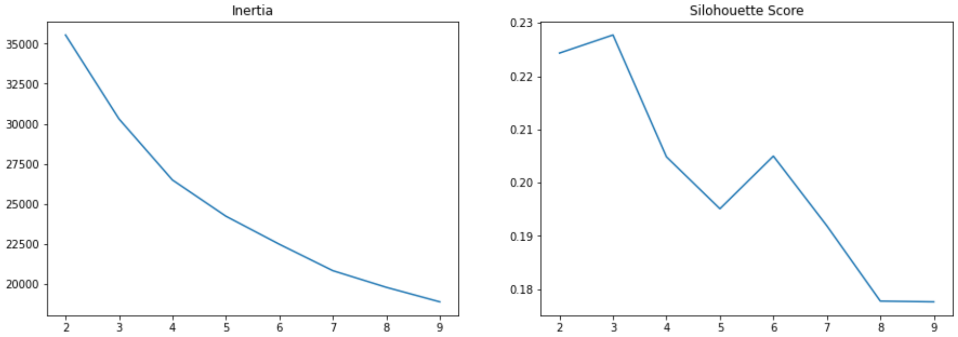
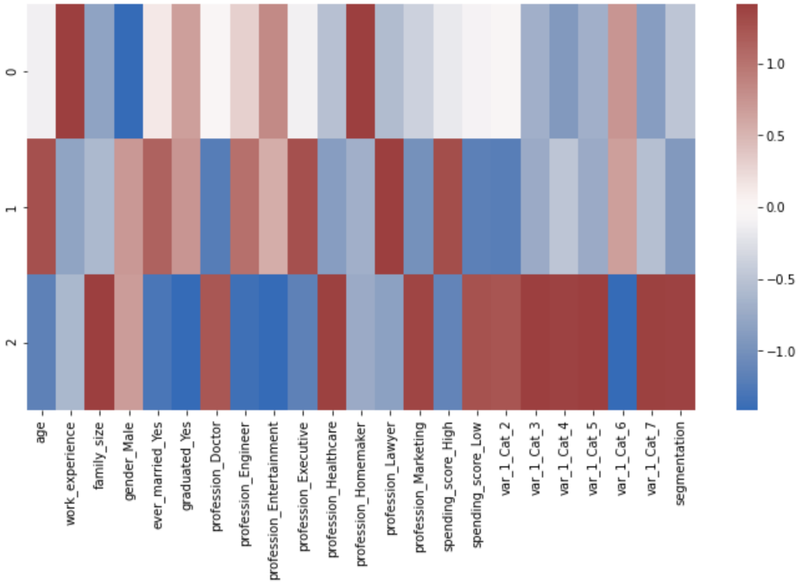


*Figure 7. Silhouette plot*



Some characteristics of the clusters can be identified through the heat map shown in Figure 7. It suggests that group 0 is highly correlated with the gender of female, and also highly correlated with work experience and the profession of homemaker. For group 1, it is highly correlated with age, married individuals whose career is related to engineering or executive, and lawyer and their spending score is high. The last group is highly correlated with people with large family size, but the individuals are neither married or graduated. These individuals mainly worked in health related fields and marketing with low spending scores.

*Figure 8. Customer characteristics heatmap.*



Based on the result, it suggests that the clustering is related to the career path of the individuals, which is related to gender. So I think the reason why they don't cluster their customers is that they are trying to protect those who may suffer from gender issues. The first group is dominated by females who are homemakers, but group 2 and group 3 are for those males who are in high paid professions and those tend to be well-known male dominated professions, such as engineer, lawyers, etc. So it would be good for the company to create a product based on such information generated from their dataset, which may be a good decision to make since they will capture the needs of their customers. However, it may actually be unethical since the algorithm is trying to make decisions based on gender, which may actually hurt those who could do better but is limited by gender.

In the future, I would like to know how a neural network would perform on this dataset since it is known for high accuracy and is suitable for both regression and classification problems. And I would also try to understand the reason why business entities make decisions based on the results from their dataset, which may help me better understand the role of data in the business world.

In our life, we actually see the power of data. And in the past year, I was trained professionally to make sense of the data. Even though there is a data ethic class,  and I was trying to take ethic into consideration when I performed data analysis, but my decision may actually hurt people, and my algorithm may decide who will benefit. It shows me the power of data analyst in designing the world. And I think that is the most important lesson I learned in this project.